Analisis Dampak *SelectKBest* dan *SMOTEENN* terhadap Akurasi Model Klasifikasi Penyakit Cacar Monyet Menggunakan Algoritma *Machine Learning*

1st Agung Triatna
Universitas Buana Perjuangan Karawang
Karawang, Indonesia
if21.agungtriatna@mhs.ubpkarawang.ac.id

2nd Yana Cahyana
Universitas Buana Perjuangan Karawang
Karawang, Indonesia
yana.cahyana@ubpkarawang.ac.id

3rd Tohirin Al Mudzakir
Universitas Buana Perjuangan Karawang
Karawang, Indonesia
tohirin@ubpkarawang.ac.id

4th Adi Rizky Pratama
Universitas Buana Perjuangan Karawang
Karawang, Indonesia
adi.rizky@ubpkarawang.ac.id

***Abstract* — Penyebaran cacar monyet yang cepat dan sulit dikendalikan membutuhkan metode prediksi penyakit yang akurat. Kesalahan prediksi *false negative* menyebabkan infeksi tidak terdeteksi. Sebaliknya, diagnosis *false positive* menimbulkan kecemasan yang tidak perlu dan membebani fasilitas kesehatan dengan kasus yang sebenarnya tidak terinfeksi. Penelitian ini dilakukan untuk mengetahui pengaruh *SelectKBest* dan *SMOTEENN* terhadap akurasi model klasifikasi penyakit cacar monyet. Dataset yang digunakan berisi rekam medis gejala klinis pasien cacar monyet dengan dimensi (25000, 11). Tahapan pengolahan data meliputi pengumpulan data, analisis data eksploratif (EDA), prapemrosesan, pemodelan, dan evaluasi. Penelitian ini menggunakan empat variasi dataset, yaitu dataset asli tanpa modifikasi, dataset hasil seleksi fitur dengan *SelectKBest*, dataset hasil resampling menggunakan *SMOTEENN*, dan dataset kombinasi *SelectKBest* dan *SMOTEENN*. Hasil penelitian menunjukkan kombinasi *SelectKBest* dan *SMOTEENN* terbukti paling efektif meningkatkan akurasi model klasifikasi. Algoritma *XGBoost* mencapai akurasi sebesar 100%, diikuti oleh *Gradient Boosting* dengan akurasi 98,57%, dan *AdaBoost* sebesar 89,97% Temuan ini menunjukkan bahwa pemilihan fitur yang tepat yang dikombinasikan dengan metode resampling data meningkatkan performa model dalam klasifikasi penyakit cacar monyet.**

Kata kunci — cacar monyet, klasifikasi, selectkbest, SMOTEENN, machine learning

I. PENDAHULUAN

Pandemi merupakan wabah penyakit yang menyebar luas ke berbagai negara atau benua, dengan penyebaran yang sulit untuk dikendalikan [1]**.** Salah satu penyakit menular yang dikategorikan sebagai pandemi oleh WHO adalah cacar monyet *(Monkeypox)*. Cacar monyet termasuk ke dalam kelompok virus *Orthopox* dari keluarga *Poxviridae* [2]. Penyakit ini dapat dengan cepat menyebar melalui kontak langsung dengan objek yang terkontaminasi.

Cacar monyet dapat menyebabkan gejala serius seperti demam tinggi, ruam kulit yang menyakitkan, dan lesi kulit yang berpotensi meninggalkan bekas luka permanen [3]. Pada individu dengan imunitas rendah, seperti anak-anak, lansia, ibu hamil dan penderita komorbiditas, penyakit ini dapat menyebabkan komplikasi serius yang berujung pada kematian. Selain dampak kesehatan, cacar monyet juga memberikan dampak sosial dan psikologis. Stigma negatif terhadap pasien dapat memicu diskriminasi, isolasi sosial, dan tekanan mental yang berat. Secara ekonomi, wabah ini dapat menambah beban finansial akibat biaya pengobatan, karantina, dan gangguan aktivitas ekonomi.

Dampak dari cacar monyet tidak hanya membebani sistem kesehatan, tetapi juga menghambat aktivitas sosial dan mengganggu stabilitas ekonomi. Hingga Maret 2025, WHO melaporkan 137.919 kasus terkonfirmasi di seluruh dunia dengan 317 kematian [4]. Di Indonesia, Kementerian Kesehatan mencatat 88 kasus terkonfirmasi hingga 17 Agustus 2024 [5]. Prediksi dini yang akurat sangat penting untuk mencegah penyebaran lebih luas dan dampak kesehatan yang lebih berat. Ketidakakuratan dalam prediksi dapat mengarah pada penanganan yang salah. Jika hasil prediksi menunjukkan negatif padahal sebenarnya positif *(false negative)* dapat menyebabkan infeksi tidak terdeteksi, memperburuk penyebaran dan keterlambatan perawatan. Sebaliknya, jika hasil prediksi menunjukkan positif padahal sebenarnya negatif *(false positive)*, hal ini dapat menimbulkan kecemasan dan membebani fasilitas kesehatan dengan kasus yang sebenarnya tidak terinfeksi.

Cacar monyet secara medis dapat dideteksi dengan tes molekuler menggunakan metode *Polymerase Chain Reaction* (PCR) [6]. Meskipun akurat, PCR memerlukan biaya tinggi untuk peralatan dan reagen, serta peralatan khusus dan keahlian teknis untuk prosedur dan interpretasi hasil, yang menjadi tantangan di laboratorium dengan sumber daya terbatas [7]. Sebagai alternatif, teknologi kecerdasan buatan telah digunakan untuk mendeteksi penyakit ini, terutama melalui metode klasifikasi, dengan berbagai algoritma pembelajaran mesin yang diterapkan untuk mengklasifikasikan gambar atau data medis yang berkaitan dengan cacar monyet.

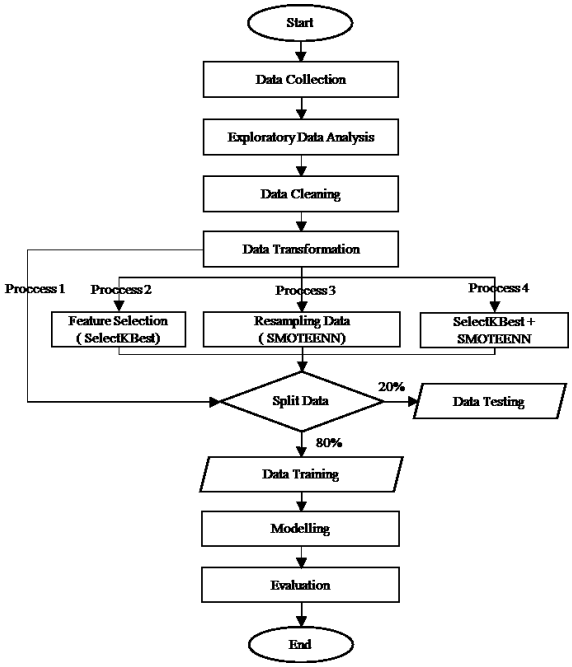
Penelitian sebelumnya telah menguji berbagai metode untuk meningkatkan akurasi dalam klasifikasi cacar monyet. Hamdan & Ekmekci (2024) menerapkan metode *voting* pada algoritma *K-Nearest Neighbors, Support Vector Classification, Random Forest, Naive Bayes*, dan *Gradient Boosting* untuk meningkatkan akurasi klasifikasi penyakit cacar monyet [8]. Hasil penelitian ini menunjukkan bahwa kombinasi algoritma *Gradient Boosting* dengan metode *voting* menghasilkan akurasi tertinggi, yaitu 63%.

Penelitian lain oleh Siena et al. (2025) menunjukkan bahwa penerapan *SMOTEENN* secara signifikan meningkatkan kinerja klasifikasi pada kasus cacar monyet dengan menggunakan algoritma *Gradient Boosting, XGBoost, dan LightGBM* [9]. Metode ini terbukti efektif dalam memperbaiki ketidakseimbangan kelas serta mengurangi noise, sehingga mampu meningkatkan akurasi semua model hingga mencapai 69%. Hasil penelitian tersebut menegaskan bahwa *SMOTEENN* merupakan metode yang efektif untuk mengatasi masalah ketidakseimbangan kelas dan noise, yang sering ditemukan pada dataset penyakit menular seperti cacar monyet.

Penelitian selanjutnya oleh Nagro (2025) menunjukan bahwa algoritma *Stacking Classifier* yang dikombinasikan dengan dataset sintesis menggunakan *CTGAN* mampu meningkatkan akurasi klasifikasi cacar monyet secara signifikan [10]. Algoritma *Stacking Classifier* menunjukkan performa terbaik dengan akurasi 87,29%, mengungguli algoritma lain seperti *LightGBM*, *XGBoost*, *LSTM*, *Tab Transformer, DenseNet-21, CNN, AdaBoost, Random Forest, Gradient Boosting, Decision Tree, SVM*, dan *Naïve Bayes*. Penggunaan data sintetis terbukti efektif dalam mengatasi ketidakseimbangan kelas, sehingga meningkatkan akurasi model.

Mengingat pentingnya model klasifikasi yang akurat untuk mencegah penyebaran lebih lanjut dan dampak kesehatan yang lebih berat, penelitian ini bertujuan untuk menganalisis pengaruh *SelectKBest* dan *SMOTEEN* terhadap akurasi model klasifikasi menggunakan algoritma *Gradient Boosting, XGBoost* dan *AdaBoost*. Diharapkan pendekatan ini dapat menghasilkan model dengan akurasi yang lebih tinggi serta memberikan kontribusi signifikan terhadap pengembangan teknik klasifikasi dalam konteks penyakit menular.

II. METODE PENELITIAN



Gambar 1 Metode Penelitian

Berdasarkan Gambar 1, penelitian ini dilaksanakan dalam delapan tahap. Dimulai dengan pengumpulan data sebagai dasar analisis, dilanjutkan dengan *Exploratory Data Analysis* (EDA) untuk memahami karakteristik data. Selanjutnya, dilakukan pembersihan data dari *missing value*, duplikat, dan *outlier*, serta transformasi data kategorikal menjadi numerik. Tahap berikutnya adalah membuat empat variasi dataset berdasarkan pendekatan yang berbeda. Variasi pertama menggunakan dataset asli tanpa modifikasi. Variasi kedua menerapkan metode *SelectKBest* untuk melakukan seleksi fitur yang paling relevan. Variasi ketiga memanfaatkan metode *SMOTEENN* untuk menangani ketidakseimbangan kelas serta mengurangi noise. Variasi keempat menerapkan kombinasi *SelectKBest* dan *SMOTEENN*. Keempat variasi dataset ini kemudian dibagi menjadi 80% untuk data latih dan 20% untuk data uji. Tahap selanjutnya membuat model klasifikasi menggunakan algoritma *Gradient Boosting, XGBoost*, dan *AdaBoost*. Terakhir, performa model dievaluasi menggunakan *Confusion Matrix* berdasarkan metrik akurasi, presisi, recall, dan f1-score.

A. Pengumpulan Data *(Collecting Data)*

Pengumpulan data adalah proses memperoleh informasi, data, atau fakta dari berbagai sumber untuk mendukung analisis, penelitian, atau pengambilan keputusan [11]. Tahapan ini sangat penting dalam pengembangan model *Machine Learning*, karena kualitas dan kuantitas data secara langsung memengaruhi kinerja serta akurasi model yang dihasilkan. Dalam penelitian ini, data diperoleh dari situs *Kaggle* dan diakses pada 30 November 2024 [12]. Dataset yang digunakan terdiri dari 25.000 baris data dan 11 kolom, yang menggambarkan berbagai gejala klinis yang dialami oleh pasien terinfeksi cacar monyet. Fitur-fitur dalam dataset ini mencakup: *Systemic Illness, Rectal Pain, Sore Throat, Penile Oedema, Oral Lesions, Solitary Lesion, Swollen Tonsils, HIV Infection, Sexually Transmitted Infection*, serta label *Monkeypox* yang menunjukkan status infeksi. Informasi lebih rinci mengenai dataset ini disajikan pada Tabel 1.

Tabel 1 Dataset Pasien Cacar Monyet

Feature	Data Type	Unique Count	Unique Values
Patient_ID	Object	25000	P0, P1, P2, …, P24997, P24998, P24999
Systemic Illness	Object	4	None, Fever, Swollen Lymph Nodes, Muscle Ache
Rectal Pain	Bool	2	False, True
Sore Throat	Bool	2	False, True
Penile Oedema	Bool	2	False, True
Oral Lesions	Bool	2	False, True
Solitary Lesion	Bool	2	False, True
Swollen Tonsils	Bool	2	False, True
HIV Infection	Bool	2	False, True
Sexually Transmitted Infection	Bool	2	False, True
MonkeyPox	Object	2	Negative, Positive

B. *Exploratory Data Analysis (EDA)*

Exploratory Data Analysis (EDA) adalah tahap awal dalam analisis data yang bertujuan untuk memahami karakteristik dan struktur dataset [13]. Proses ini melibatkan penggunaan teknik statistik deskriptif dan visualisasi untuk mengeksplorasi distribusi data, mendeteksi anomali, serta mengidentifikasi hubungan antar variabel. EDA membantu peneliti dalam memeriksa asumsi dasar, mengajukan hipotesis awal tentang fitur yang mungkin relevan, dan mengidentifikasi potensi masalah sebelum tahap pemodelan. Aspek yang dianalisis meliputi statistik deskriptif, visualisasi (seperti histogram dan scatter plot), analisis korelasi, serta deteksi outlier dan anomali data.

C. *Data Preprocessing*

Preprocessing adalah serangkaian langkah yang dilakukan untuk mempersiapkan dataset agar siap digunakan dalam proses pemodelan [14]. Tahap ini mencakup berbagai proses seperti pembersihan data *(data cleaning)*, normalisasi atau standarisasi, penanganan *missing value*, *transformasi data*, seleksi fitur, serta resampling data. Pada penelitian ini, tahap *preprocessing* yang akan dilakukan antara lain :

1. Transformasi Data *(Data Transformation)*

Data Transformation adalah proses mengubah dataset menjadi format yang dapat digunakan untuk analisis dan pemodelan *Machine Learning* [15]. Pada tahap ini, dilakukan proses *encoding* untuk mengonversi data kategorikal menjadi bentuk numerik agar dapat digunakan dalam pelatihan model klasifikasi. Dalam penelitian ini, dua metode transformasi data yang akan digunakan adalah *Label Encoding* dan *One-Hot Encoding*. *Label Encoding* adalah metode yang digunakan untuk mengubah data kategorikal menjadi numerik, di mana setiap kategori unik diberikan nilai integer [16]. Metode ini cocok digunakan ketika data kategorikal memiliki urutan atau ordinalitas, seperti peringkat atau tingkat, di mana perbedaan antar kategori dapat diukur. *One-Hot Encoding* adalah metode yang mengubah data kategorikal menjadi representasi biner dengan membuat kolom baru untuk setiap kategori unik. Dalam teknik ini, hanya satu kolom yang bernilai 1 (untuk kategori yang relevan), sementara kolom lainnya bernilai 0. Metode ini lebih tepat digunakan ketika data kategorikal tidak memiliki urutan atau ordinalitas.

2. Seleksi Fitur *(Feature Selection)*

Seleksi fitur adalah proses memilih fitur paling relevan dan berpengaruh terhadap target yang ingin diprediksi dalam sebuah model *Machine Learning* [17]. Pada penelitian ini, metode seleksi fitur yang diterapkan adalah *SelectKBest*, yang berfungsi untuk memilih fitur-fitur terbaik dari dataset berdasarkan hubungan korelasi dengan variabel target [18]. Metode ini bekerja dengan menghitung skor relevansi antara setiap fitur dan target, kemudian memilih K fitur teratas dengan skor tertinggi untuk digunakan dalam proses pelatihan model. Tujuan dari metode ini adalah untuk meningkatkan akurasi model dengan mengeliminasi fitur-fitur yang tidak relevan atau kurang berpengaruh terhadap target. Skor setiap fitur terhadap target dihitung menggunakan Persamaan (1), dan fitur-fitur dengan skor tertinggi dipilih hingga mencapai jumlah fitur yang telah ditentukan (K).

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

Keterangan :

O_i = Frekuensi yang diamati

E_i = Frekuensi yang diharapkan

3. Data Resampling

Resampling data merupakan teknik statistik yang digunakan untuk membuat sampel data baru dari satu sampel data yang sudah ada [19]. Tujuan utama dari *resampling* adalah untuk mengatasi masalah ketidakseimbangan kelas yang dapat menurunkan kinerja model *Machine Learning*, terutama dalam mengenali kelas minoritas. Dalam penelitian ini, teknik resampling yang digunakan adalah *SMOTEENN*, yang merupakan gabungan dari *SMOTE (Synthetic Minority Over-sampling Technique)* dan *ENN (Edited Nearest Neighbors)*. *SMOTE* diterapkan untuk menghasilkan data sintetis dari kelas minoritas dengan tujuan untuk menyeimbangkan jumlah data antara setiap kelas [19]. Sementara itu, *ENN* berfungsi untuk meningkatkan kualitas data dengan menghapus data yang dianggap sebagai noise atau hasil klasifikasi yang salah berdasarkan algoritma *k-Nearest Neighbors (k-NN)* [20]. Kombinasi kedua teknik ini menghasilkan dataset yang tidak hanya lebih seimbang secara kuantitatif, tetapi juga lebih bersih dan representatif.

D. Pembuatan Model Klasifikasi *(Classification Modelling)*

Classification Modelling adalah proses melatih model komputer menggunakan teknik *supervised learning* untuk mengenali pola dalam data, sehingga model dapat memprediksi kelas atau kategori pada data baru [21]. Dalam penelitian ini, tiga algoritma yang digunakan adalah *Gradient Boosting*, *XGBoost* dan *AdaBoost*.

1. Gradient Boosting

Gradient Boosting merupakan algoritma pembelajaran mesin yang tergolong dalam metode *ensemble* [9]. Algoritma ini berfungsi dengan cara membangun model secara berurutan, di mana setiap model baru yang ditambahkan bertujuan untuk mengoreksi kesalahan yang ada pada model sebelumnya. Model yang digunakan biasanya berupa pohon keputusan sederhana *(weak learners)*. Proses ini dilakukan dengan meminimalkan *error* menggunakan pendekatan *gradien* dari fungsi *loss*, sehingga disebut *Gradient Boosting*. Dengan menggabungkan banyak model lemah, *Gradient Boosting* mampu membentuk model yang kuat dan menghasilkan prediksi yang akurat.

2. Extreme Gradient Boosting (XGBoost)

XGBoost adalah algoritma *Machine Learning* berbasis pohon keputusan yang bekerja dengan membangun serangkaian pohon secara bertahap [22]. Setiap pohon yang dibangun bertujuan untuk memperbaiki kesalahan prediksi yang dihasilkan oleh pohon sebelumnya. Salah satu keunggulan *XGBoost* dibandingkan algoritma boosting lainnya adalah diperkenalkannya konsep regularisasi ke dalam fungsi objektif, yang berfungsi untuk mengurangi risiko *overfitting* [23]. Fungsi objektif dari *XGBoost* dijelaskan dalam Persamaan (2).

$$O = \sum_{i=1}^{n} L\left(yi, F(xi)\right) + \sum_{k=1}^{t} R(fk) + C \quad (2)$$

Keterangan:

$L\left(yi, F(xi)\right)$: fungsi loss antara nilai aktual dan prediksi

$R(fk)$: fungsi regularisasi untuk kompleksitas model

C : Konstanta

3. Adaptive Boosting (AdaBoost)

Algoritma *AdaBoost* adalah metode *ensemble learning* yang menggabungkan beberapa *weak learners* secara sekuensial dengan memberi bobot lebih besar pada sampel yang salah klasifikasi untuk membentuk *strong learner* yang lebih akurat [24]. Mekanisme adaptif ini memungkinkan model untuk fokus secara bertahap pada data sulit dengan memperbarui bobot sampel di setiap iterasi, meskipun rentan terhadap *outlier* dan *noisy* data, sehingga pemilihan jumlah iterasi dan kontrol *learning rate* menjadi sangat penting agar tidak terjadi *overfitting*.

E. *Evaluasi Model*

Evaluasi model adalah proses untuk mengukur dan menilai kinerja suatu model *Machine Learning* berdasarkan data uji yang tidak digunakan selama proses pelatihan [25]. Tujuan utama dari evaluasi ini adalah untuk mengetahui seberapa baik model mampu melakukan prediksi terhadap data baru, serta untuk membandingkan efektivitas antar model yang berbeda. Pada penelitian ini, evaluasi model klasifikasi akan memanfaatkan *Confusion Matrix* sebagai alat analisis performa. *Confusion Matrix* memberikan representasi visual mengenai kemampuan model dalam membedakan kelas target berdasarkan prediksi dan nilai aktual. Matriks ini mencakup empat elemen utama: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Berdasarkan *Confusion Matrix*, metrik evaluasi seperti Akurasi, Presisi, Recall, dan F1-Score dapat dihitung, dengan rumus yang dijelaskan pada Persamaan (3), (4), (5), dan (6).

$$Accuracy = \frac{True\ Positive+True\ Negative}{SUM\ The\ Number\ of\ Data}\ x\ 100\% \quad (3)$$

$$Precision = \frac{True\ Positive}{True\ Positive+False\ Positive}\ x\ 100\% \quad (4)$$

$$Recall = \frac{True\ Positive}{True\ Positive+False\ Negative}\ x\ 100\% \quad (5)$$

$$F1 - Score = 2\ x\ \frac{precision\ x\ recall}{precision\ +recall}\ x\ 100\% \quad (6)$$

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data *(Collecting Data)*

Langkah awal dalam penelitian ini adalah melakukan persiapan data yang akan digunakan dalam analisis. Proses ini diawali dengan mengimpor sejumlah *library* penting yang mendukung berbagai aktivitas seperti pengolahan data, pembuatan visualisasi, serta pengembangan model *Machine Learning*. Setelah semua *library* berhasil dimuat, langkah selanjutnya adalah mengakses dataset yang disimpan di *Google Drive*, yang kemudian dibaca ke dalam program.

Setelah dataset berhasil dimuat, langkah selanjutnya adalah menampilkan isi data untuk melakukan peninjauan awal. Pada tahap ini, dilakukan identifikasi terhadap jumlah kolom dan baris, jenis data dari masing-masing fitur, serta informasi dasar lainnya yang akan menjadi acuan dalam proses eksplorasi dan analisis data selanjutnya. Untuk detailnya dapat dilihat pada Gambar 2.

	Patient_ID	Systemic Illness	Rectal Pain	Sore Throat	Penile Oedema	Oral Lesions	Solitary Lesion	Swollen Tonsils	HIV Infection	Sexually Transmitted Infection	MonkeyPox
0	P0	None	False	True	True	True	False	True	False	False	Negative
1	P1	Fever	True	False	True	True	False	False	True	False	Positive
2	P2	Fever	False	True	True	False	False	False	True	False	Positive
3	P3	None	True	False	False	False	True	True	True	False	Positive
4	P4	Swollen Lymph Nodes	True	True	True	False	False	True	True	False	Positive
...	...	...	...	...	...	...	...	...	...	...	...
24995	P24995	None	True	True	False	True	True	False	False	True	Positive
24996	P24996	Fever	False	True	True	False	True	True	True	True	Positive
24997	P24997	None	True	True	False	False	True	True	False	False	Positive
24998	P24998	Swollen Lymph Nodes	False	True	False	True	True	True	False	False	Negative
24999	P24999	Swollen Lymph Nodes	False	False	True	False	False	True	True	False	Positive

25000 rows × 11 columns

Gambar 2 Membaca dan Menampilkan Dataset

B. *Exploratory Data Analysis (EDA)*

Dataset yang telah dikumpulkan kemudian dianalisis lebih lanjut untuk memperoleh wawasan yang lebih dalam. Dari hasil eksplorasi, diketahui bahwa dataset ini berisi 25.000 entri data dengan total 11 fitur yang menggambarkan gejala pasien yang terinfeksi *Monkeypox*. Dataset ini mencakup delapan fitur bertipe data *boolean*, yaitu: *Rectal Pain, Sore Throat, Penile Oedema, Oral Lesions, Solitary Lesion, Swollen Tonsils, HIV Infection,* serta *Sexually Transmitted Infection*. Setiap fitur ini bernilai *true* atau *false*, yang menunjukkan apakah gejala atau kondisi tersebut ada pada pasien.

Selain itu, terdapat tiga fitur yang memiliki tipe data objek, yaitu *Patient_ID, Systemic Illness,* dan *Monkeypox*. Fitur *Patient_ID* berisi nomor identifikasi pasien secara berurutan dan dikategorikan sebagai data ordinal karena nilainya dapat diurutkan berdasarkan urutan pasien. Fitur *Systemic Illness* menggambarkan kondisi klinis yang dialami oleh pasien dan termasuk dalam tipe data nominal, karena menunjukkan kategori penyakit sistemik yang tidak memiliki urutan hierarkis. Adapun fitur *Monkeypox* merupakan label target klasifikasi, yang menunjukkan apakah pasien positif atau negatif terhadap infeksi *Monkeypox*. Untuk detailnya dapat dilihat pada Gambar 3.

```
original_dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype 
---  ------                          --------------  ----- 
 0   Patient_ID                      25000 non-null  object
 1   Systemic Illness                25000 non-null  object
 2   Rectal Pain                     25000 non-null  bool  
 3   Sore Throat                     25000 non-null  bool  
 4   Penile Oedema                   25000 non-null  bool  
 5   Oral Lesions                    25000 non-null  bool  
 6   Solitary Lesion                 25000 non-null  bool  
 7   Swollen Tonsils                 25000 non-null  bool  
 8   HIV Infection                   25000 non-null  bool  
 9   Sexually Transmitted Infection  25000 non-null  bool  
 10  MonkeyPox                       25000 non-null  object
dtypes: bool(8), object(3)
memory usage: 781.4+ KB
```

Gambar 3 Informasi Struktur Dataset Pasien Cacar Monyet

ISSN: 2715-2766
Vol. 7 No. 1, Januari 2026

Hasil analisa lebih lanjut ditemukan *missing value* sebanyak 6.216 pada fitur *Systemic Illness*, detailnya dapat dilihat pada Gambar 4 (a). Selain itu terdapat ketidakseimbangan pada kelas target, di mana jumlah data dengan label *Positive* mencapai 15.909, sementara data dengan label *Negative* hanya 9.091 detailnya dapat dilihat pada Gambar 4 (b). Hasil analisis korelasi juga menunjukkan bahwa tidak semua fitur memiliki hubungan yang kuat dengan target, detailnya dapat dilihat pada Gambar 4 (c). Selain itu, ditemukan ketidakkonsistenan dalam kelas target, yaitu adanya kasus dengan nilai fitur yang identik tetapi memiliki label yang berbeda, detailnya dapat dilihat pada Tabel 3. Dengan demikian, dapat disimpulkan bahwa terdapat empat permasalahan utama yang berpotensi memengaruhi akurasi model klasifikasi, yaitu: (1) nilai hilang *(missing value)*, (2) ketidakseimbangan distribusi kelas pada variabel target *(class imbalance)*, (3) inkonsistensi dalam pelabelan target *(label noise)*, dan (4) rendahnya korelasi antara sebagian fitur dengan variabel target.

(a) (b) (c)

Gambar 4 Missing Value (a), Distribusi Target (b) dan Korelasi Antar Fitur (c)

Tabel 2 Anomali Dataset Pasien Cacar Monyet

Patient_ID	Systemic Illness	Rectal Pain	Sore Throat	Penile Oedema	Oral Lesions	Solitary Lesion	Swollen Tonsils	HIV Infection	Sexually Transmitted Infection	MonkeyPox
P1684	Fever	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Negative
P3061	Fever	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Positive
P3883	Fever	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Negative
P4395	Fever	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Negative
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
P1970	Swollen Lymph Nodes	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Positive
P6787	Swollen Lymph Nodes	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Positive
P7330	Swollen Lymph Nodes	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Negative

C. *Preprosesing*

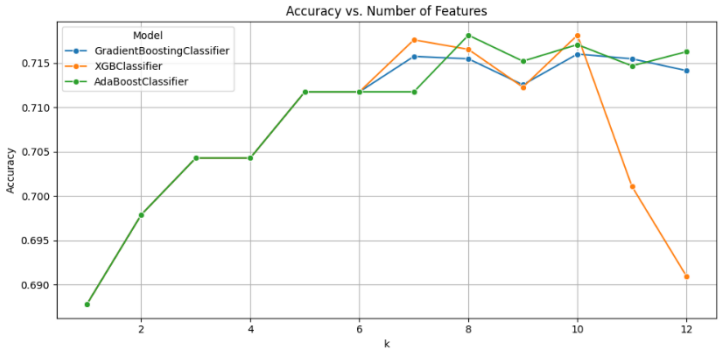
Tahap *preprocessing* ini dimulai dengan proses transformasi data, yang kemudian dilanjutkan dengan pembuatan empat variasi dataset. Variasi pertama berupa data asli tanpa modifikasi, variasi kedua menggunakan metode *SelectKBest* untuk seleksi fitur, variasi ketiga menerapkan teknik resampling menggunakan *SMOTEENN*, dan variasi keempat menerapkan kombinasi dari *SelectKBest* dan *SMOTEENN*.

1. Dataset Asli Tanpa Modifikasi

Metode *preprocessing* yang diterapkan pada variasi dataset ini mencakup *data cleaning* dan *data transformation*. Sebelum pembersihan, dataset memiliki 6.216 *missing value* pada fitur *Systemic Illness* dengan dimensi awal (25.000, 11). Setelah baris yang mengandung *missing value* dihapus, jumlah data berkurang menjadi 18.784, sehingga dimensi menjadi (18.784, 11). Selanjutnya, dilakukan *data transformation* dengan metode *Label Encoding* untuk fitur *Patient_ID* dan *MonkeyPox* yang bersifat ordinal, serta *One Hot Encoding* untuk fitur *Systemic Illness* yang bersifat nominal. Transformasi ini menambah jumlah fitur, sehingga dimensi akhir menjadi (18.784, 13).

2. Dataset dengan Seleksi Fitur *SelectKBest*

Pada varian ini, dataset diproses melalui tiga tahap, yaitu *data cleaning, data transformation,* dan seleksi fitur. Tahap pertama, *data cleaning,* dilakukan dengan menghapus baris yang mengandung *missing value,* sehingga dimensi dataset berubah dari (25.000, 11) menjadi (18.784, 11). Setelah itu, proses *data transformation* diterapkan untuk mengubah fitur kategorial menjadi numerik yang sesuai dengan kebutuhan algoritma klasifikasi. Hasil dari transformasi ini menyebabkan perubahan dimensi dataset menjadi (18.784, 13). Selanjutnya, dilakukan seleksi fitur menggunakan metode *SelectKBest* dengan mencoba nilai K dari 1 hingga 12 untuk menentukan jumlah fitur terbaik yang dapat meningkatkan performa masing-masing algoritma klasifikasi. Berdasarkan hasil evaluasi, diperoleh nilai K yang optimal untuk setiap algoritma, yakni K = 10 untuk *Gradient Boosting*, K = 10 untuk *XGBoost*, dan K = 8 untuk *AdaBoost*. Untuk algoritma *Gradient Boosting* dan *XGBoost*, fitur terbaik yang terpilih adalah *Rectal Pain, Sore Throat, Penile Oedema, Oral Lesions, Solitary Lesion, HIV Infection, Sexually Transmitted Infection, Systemic Illness_Fever, Systemic Illness_Muscle Aches and Pain, Systemic Illness_Swollen Lymph Nodes,* dengan *MonkeyPox* sebagai label target. Sementara itu, untuk algoritma *AdaBoost*, fitur terbaik yang terpilih adalah *Rectal Pain, Sore Throat, Penile Oedema, HIV Infection, Sexually Transmitted Infection, Systemic Illness_Fever, Systemic Illness_Muscle Aches and Pain, dan Systemic Illness_Swollen Lymph Nodes,* dengan *MonkeyPox* sebagai label target. Hasil seleksi fitur serta evaluasi kinerja masing-masing algoritma terhadap variasi nilai K dapat dilihat pada Gambar 5.



Gambar 5 Grafik Perbandingan Akurasi Terhadap Jumlah Fitur (K)

3. Dataset dengan Resampling SMOTEENN

Pada varian ini, proses pengolahan data terdiri atas tiga tahapan utama, yaitu *data cleaning, data transformation*, dan *resampling* menggunakan metode *SMOTEENN*. Tahapan data *cleaning* dan *data transformation* dilakukan dengan prosedur yang identik dengan varian sebelumnya. Perbedaan utama terletak pada tahap akhir, yakni penerapan metode *resampling* untuk mengatasi permasalahan ketidakseimbangan distribusi kelas serta mengurangi data yang tidak konsisten. Dataset semula berdimensi (25.000, 11). Setelah dilakukan data *cleaning*, sebanyak 6.216 entri yang mengandung missing value dihapus, sehingga jumlah data tersisa menjadi 18.784 baris dengan 11 fitur. Selanjutnya, pada tahap *data transformation*, dilakukan penambahan dua fitur baru hasil proses transformasi, yang menyebabkan dimensi dataset berubah menjadi (18.784, 13). Tahapan *resampling* dilakukan dengan menerapkan metode *SMOTEENN* yang mengombinasikan pendekatan *oversampling* dan *undersampling*. Pada tahap pertama, SMOTE digunakan untuk melakukan penyeimbangan distribusi kelas melalui sintesis data pada kelas minoritas. Setelah distribusi kelas menjadi seimbang, tahap berikutnya dilakukan penyaringan data menggunakan metode *Edited Nearest Neighbours* (ENN), yang bertujuan untuk mengeliminasi data anomali. Melalui kombinasi dua teknik ini, diperoleh dataset akhir dengan dimensi (7.368, 13).

4. Dataset dengan Kombinasi SelectKBest dan SMOTEENN

Pada varian ini, dataset diproses melalui beberapa tahapan, yaitu *data cleaning*, *data transformation*, seleksi fitur menggunakan *SelectKBest*, dan diakhiri dengan proses *resampling* menggunakan metode *SMOTEENN*. Setiap tahapan memberikan dampak terhadap perubahan dimensi dataset yang digunakan untuk melatih masing-masing algoritma klasifikasi.

Untuk algoritma *Gradient Boosting*, dataset semula berdimensi (25.000, 11). Setelah dilakukan *data cleaning*, sebanyak 6.216 baris data dengan *missing value* dihapus sehingga dimensi menjadi (18.784, 11). Selanjutnya, proses *data transformation* menambahkan dua fitur baru, menjadikan dimensi berubah menjadi (18.784, 13). Setelah dilakukan seleksi fitur dengan *SelectKBest*, jumlah fitur diseleksi kembali menjadi 11, menghasilkan dimensi (18.784, 11). Terakhir, penerapan metode *SMOTEENN* mengubah dimensi menjadi (6.663, 11).

Pada dataset yang digunakan untuk melatih algoritma *XGBoost*, proses perubahan dimensi identik dengan yang terjadi pada *Gradient Boosting*. Dimensi awal dataset adalah (25.000, 11). Setelah dilakukan *data cleaning*, sebanyak 6.216 baris data yang mengandung *missing value* dihapus, sehingga dimensi dataset menjadi (18.784, 11). Selanjutnya, pada tahap *data transformation*, dua fitur tambahan dihasilkan, yang mengubah dimensi menjadi (18.784, 13). Setelah dilakukan seleksi fitur menggunakan metode *SelectKBest*, jumlah fitur dikurangi kembali menjadi 11, sehingga dimensi menjadi (18.784, 11). Terakhir, penerapan metode *SMOTEENN* menghasilkan dataset akhir dengan dimensi (6.663, 11).

Sementara itu, dataset yang digunakan untuk melatih algoritma *AdaBoost* mengalami pola perubahan yang hampir serupa, dengan perbedaan pada hasil akhir seleksi fitur. Dataset awal berdimensi (25.000, 11), kemudian menjadi (18.784, 11) setelah *data cleaning*, dan meningkat menjadi (18.784, 13) setelah *data transformation.* Setelah dilakukan seleksi fitur menggunakan *SelectKBest*, jumlah fitur terpilih berkurang menjadi 9, sehingga dimensi menjadi (18.784, 9). Setelah melalui proses resampling dengan *SMOTEENN*, dimensi dataset akhir menjadi (6.478, 9).

D. Training Model

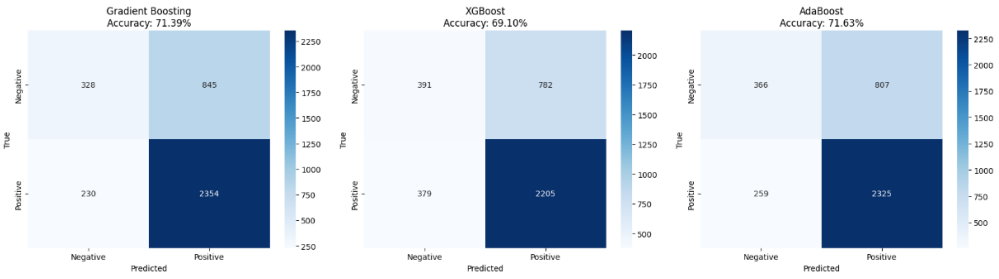
Pada tahap ini, dilakukan percobaan pelatihan model klasifikasi dengan menggunakan empat varian dataset untuk mengevaluasi pengaruh metode *SelectKBest* dan *SMOTEENN* terhadap performa model. Pembagian data dilakukan dengan rasio 80% untuk data pelatihan *(training)* dan 20% untuk data pengujian *(testing),* menggunakan parameter random_state = 42 untuk memastikan hasil yang konsisten dan dapat direproduksi. Varian pertama merupakan dataset asli yang tidak mengalami modifikasi. Varian kedua menggunakan teknik *SelectKBest* untuk melakukan seleksi fitur dengan memilih fitur-fitur paling relevan berdasarkan tingkat korelasi terhadap target. Varian ketiga menerapkan metode *SMOTEENN* yang bertujuan untuk menyeimbangkan distribusi kelas serta menghapus data yang tidak konsisten. Sementara itu, varian keempat merupakan kombinasi dari dua metode sebelumnya, dengan urutan proses berupa seleksi fitur menggunakan *SelectKBest* terlebih dahulu, kemudian diikuti oleh resampling menggunakan *SMOTEENN* pada fitur-fitur yang telah terpilih. Tiga algoritma pembelajaran mesin yang digunakan dalam proses pelatihan model adalah *Gradient Boosting, XGBoost,* dan *AdaBoost*. Setelah proses pelatihan selesai, performa setiap kombinasi varian dataset dan algoritma dibandingkan menggunakan *Confusion Matrix*, dengan mengukur metrik *Accuracy, Precision, Recall,* dan *F1 Score.*

E. *Evaluasi Model*

Pada tahap evaluasi, performa ketiga model *(Gradient Boosting, XGBoost,* dan *AdaBoost)* diukur dengan menggunakan *Confusion Matrix* pada empat varian dataset. Evaluasi pertama dilakukan terhadap model yang dilatih menggunakan dataset asli tanpa modifikasi. Berdasarkan hasil evaluasi ini, akurasi tertinggi diperoleh oleh algoritma *AdaBoost* sebesar 71,63%, diikuti oleh *Gradient Boosting* sebesar 71,39%, dan *XGBoost* sebesar 69,10%. Rincian lebih lanjut mengenai hasil evaluasi ini disajikan pada Tabel 3 dan Gambar 6.

Tabel 3 Hasil Evaluasi Model Klasifikasi dengan Dataset Asli Tanpa Modifikasi

Model	Accuracy	Recall	Precision	F1 Score
Gradient Boosting	71.39%	71.39%	68.96%	67.82%
XGBoost	69.10%	69.10%	66.62%	67.01%
AdaBoost	71.63%	71.63%	69.34%	68.66%

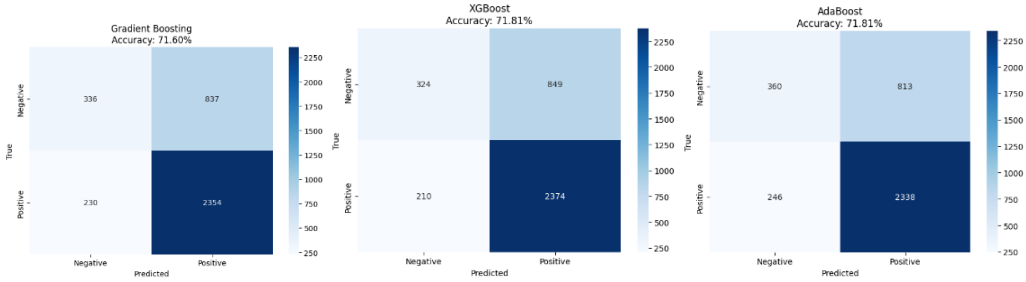


Gambar 6 Confusion Matrix Model Klasifikasi dengan Dataset Asli Tanpa Modifikasi

Evaluasi kedua dilakukan terhadap model yang dilatih menggunakan dataset hasil seleksi fitur. Hasil evaluasi menunjukkan bahwa penerapan metode *SelectKBest* mampu meningkatkan akurasi pada ketiga algoritma yang diuji. Akurasi algoritma *Gradient Boosting* mengalami peningkatan sebesar 0,21%, dari 71,39% menjadi 71,60%. Sementara itu, *XGBoost* menunjukkan peningkatan akurasi yang paling signifikan, yakni sebesar 2,71%, dari 69,10% menjadi 71,81%. Algoritma *AdaBoost* juga mengalami peningkatan sebesar 0,18%, dari 71,63% menjadi 71,81%. Rincian hasil evaluasi secara lengkap dapat dilihat pada Tabel 4 dan Gambar 7.

Tabel 4 Hasil Evaluasi Model Klasifikasi Pada Dataset Setelah Seleksi Fitur

Model	Accuracy	Recall	Precision	F1 Score
Gradient Boosting dengan SelectKBest	71.60%	71.60%	69.27%	68.13%
XGBoost dengan SelectKBest	71.81%	71.81%	69.60%	68.08%
AdaBoost dengan SelectKBest	71.81%	71.81%	69.58%	68.71%

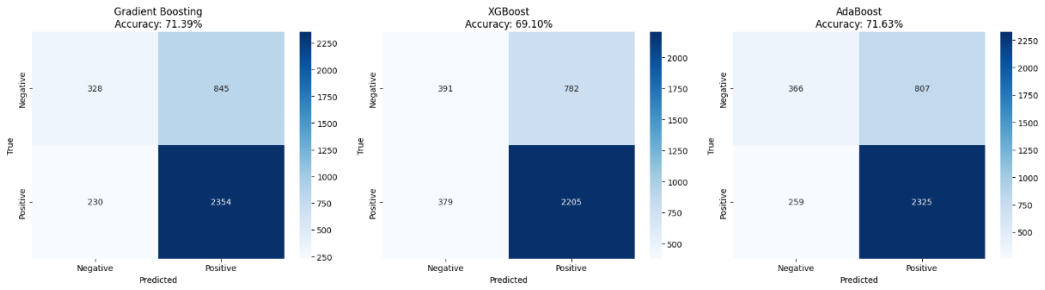


Gambar 7 *Confusion Matrix* Model Pada Dataset Hasil Seleksi Fitur

Evaluasi ketiga dilakukan pada model yang dilatih menggunakan dataset hasil *resampling* dengan metode *SMOTEENN*, sebagaimana ditunjukkan pada Gambar 4.10. Hasil evaluasi menunjukkan bahwa model yang dilatih dengan data hasil *resampling* ini menghasilkan akurasi yang lebih tinggi dibandingkan dengan model yang menggunakan dataset asli maupun dataset hasil seleksi fitur. Algoritma *XGBoost* mencatat akurasi tertinggi sebesar 80,26%, diikuti oleh algoritma *Gradient Boosting* dengan akurasi sebesar 79,78%, serta *AdaBoost* yang memperoleh akurasi sebesar 78,83%. Rincian hasil evaluasi secara lengkap disajikan pada Tabel 5 dan Gambar 8.

Tabel 5 Hasil Evaluasi Model Klasifikasi Pada Dataset Setelah *Resampling*

Model	Accuracy	Recall	Precision	F1 Score
Gradient Boosting dengan SMOTEENN	79.78%	79.78%	79.82%	79.80%
XGBoost dengan SMOTEENN	80.26%	80.26%	80.44%	80.33%
AdaBoost dengan SMOTEENN	78.83%	78.83%	78.49%	78.58%

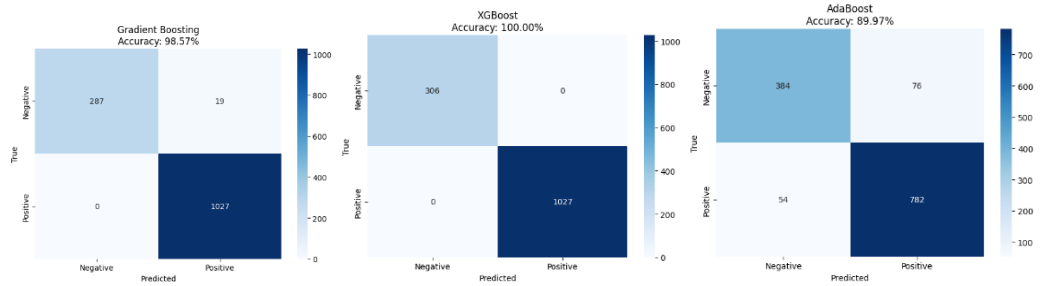


Gambar 8 *Confusion Matrix* Model Pada Dataset Hasil *Resampling*

Evaluasi keempat dilakukan pada model yang dilatih menggunakan dataset hasil kombinasi metode *SelectKBest* dan *SMOTEENN*. Hasil evaluasi menunjukkan bahwa akurasi model pada varian ini lebih tinggi dibandingkan dengan model yang dilatih menggunakan dataset varian lain. Berdasarkan hasil yang diperoleh, algoritma *XGBoost* mencapai akurasi sebesar 100%, diikuti oleh *Gradient Boosting* dengan akurasi 98,57%, dan *AdaBoost* sebesar 89,97%. Rincian lengkap hasil evaluasi disajikan pada Tabel 6 dan Gambar 9.

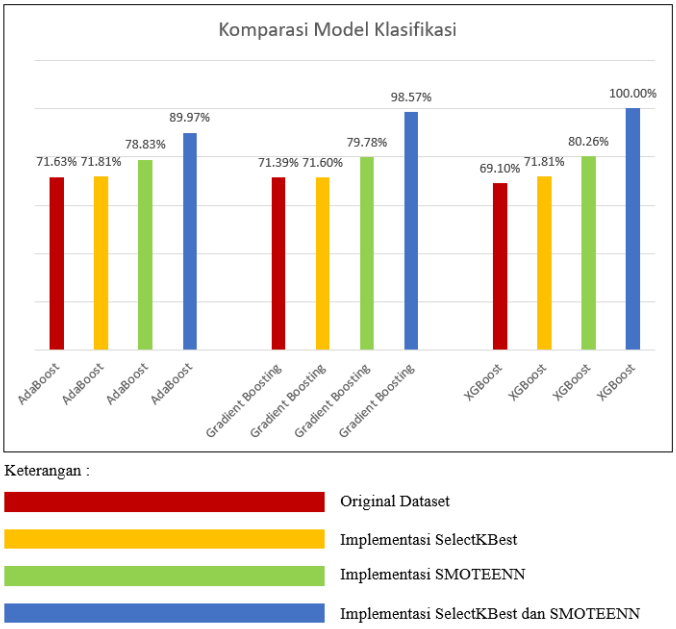
Tabel 6 Hasil Evaluasi Model Klasifikasi Pada Dataset Hasil Kombinasi *SelectKBest* dan *SMOTEENN*

Model	Accuracy	Recall	Precision	F1 Score
Gradient Boosting dengan SelectKBest & SMOTEENN	98.57%	98.57%	98.60%	98.60%
XGBoost dengan SelectKBest & SMOTEENN	100.00%	100.00%	100.00%	100.00%
AdaBoost dengan SelectKBest & SMOTEENN	89.97%	89.97%	89.91%	89.91%



Gambar 9 *Confusion Matrix* Model Pada Dataset Hasil Kombinasi *SelectKBest* dan *SMOTEENN*

Berdasarkan evaluasi terhadap empat model klasifikasi penyakit cacar monyet menggunakan algoritma *Gradient Boosting, XGBoost*, dan *AdaBoost*, diperoleh hasil bahwa kombinasi antara teknik seleksi fitur *SelectKBest* dan metode *resampling SMOTEENN* memberikan kontribusi paling signifikan terhadap peningkatan akurasi model. Penggunaan *SMOTEENN* secara mandiri menempati peringkat kedua dalam hal pengaruh terhadap akurasi, sedangkan penerapan *SelectKBest* saja menunjukkan dampak yang paling rendah dibandingkan metode lain. Untuk perbadingan tingkat akurasi setiap model dapat dilihat pada Gambar 10.



Gambar 10 Komparasi Model Klasifikasi

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan bahwa penerapan metode *SelectKBest* mampu meningkatkan akurasi pada ketiga algoritma klasifikasi yang diuji *(Gradient Boosting, XGBoost,AdaBoost*) dengan peningkatan tertinggi sebesar 2,71% pada *XGBoost*. Selain itu, penggunaan teknik resampling *(SMOTEENN)* untuk mengatasi ketidakseimbangan kelas dan ketidakkonsistenan label juga berkontribusi signifikan terhadap peningkatan akurasi, di mana *XGBoost* mencatat akurasi tertinggi sebesar 80,26%. Kombinasi antara *SelectKBest* dan *SMOTEENN* menghasilkan performa terbaik, dengan akurasi mencapai 100% pada *XGBoost*, 98,57% pada *Gradient Boosting*, dan 89,97% pada *AdaBoost*. Temuan ini menunjukkan bahwa kombinasi *SelectKBest* dengan *SMOTEENN* merupakan pendekatan yang paling efektif dalam meningkatkan kinerja model klasifikasi pada dataset penyakit cacar monyet.

B. Saran

Berdasarkan hasil dan temuan dalam penelitian ini, terdapat beberapa saran yang dapat dijadikan pertimbangan untuk penelitian selanjutnya. Pertama, mengingat ditemukannya ketidakkonsistenan label target pada tahap E*xploratory Data Analysis* (EDA), disarankan agar seluruh dataset divalidasi ulang oleh tenaga medis untuk meningkatkan kualitas data dan menghindari pembelajaran model dari informasi yang tidak akurat. Kedua, disarankan untuk mengeksplorasi metode lain dalam mengatasi ketidakkonsistenan label target, seperti penerapan *CleanLab* yang tersedia dalam *Library Scikit-learn*, untuk memperoleh hasil klasifikasi yang lebih baik.

REFERENCES

[1] N. Mascie Taylor and K. Moji, “Pandemics,” *J. Peace Nucl. Disarm.*, vol. 4, no. sup1, pp. 47–59, Mar. 2021, doi: 10.1080/25751654.2021.1880769.

[2] H. Harapan *et al.*, “Monkeypox: A Comprehensive Review,” *Viruses*, vol. 14, no. 10, p. 2155, Sep. 2022, doi: 10.3390/v14102155.

[3] J. P. Thornhill *et al.*, “Monkeypox Virus Infection in Humans across 16 Countries — April–June 2022,” *N. Engl. J. Med.*, vol. 387, no. 8, pp. 679–691, Aug. 2022, doi: 10.1056/NEJMoa2207323.

[4] World Health Organization, “Mpox,” https://www.who.int/news-room/fact-sheets/detail/mpox.

[5] M. E. dr. Siti Nadia Tarmizi and K. K. R. Biro Komunikasi dan Pelayanan Publik, "88 Kasus Konfirmasi Mpox di Indonesia, Seksual Sesama Jenis Jadi Salah Satu Penyebab." Accessed: Nov. 17, 2024. [Online]. Available: https://kemkes.go.id/id/88-kasus-konfirmasi-mpox-di-indonesia-seksual-sesama-jenis-jadi-salah-satu-penyebab

[6] V. De Pace *et al.*, "Molecular Diagnosis of Human Monkeypox Virus during 2022–23 Outbreak: Preliminary Evaluation of Novel Real-Time Qualitative PCR Assays," *Microorganisms*, vol. 12, no. 4, p. 664, Mar. 2024, doi: 10.3390/microorganisms12040664.

[7] Z. L. Chelsky, D. Dittmann, T. Blanke, M. Chang, E. Vormittag-Nocito, and L. J. Jennings, "Validation Study of a Direct Real-Time PCR Protocol for Detection of Monkeypox Virus," *J. Mol. Diagnostics*, vol. 24, no. 11, pp. 1155–1159, Nov. 2022, doi: 10.1016/j.jmoldx.2022.09.001.

[8] A. Hamdan and D. Ekmekci, "Design of Monkeypox Disease Diagnosis Model Using Classical Machine Learning Algorithm," *J. Soft Comput. Artif. Intell.*, vol. 5, no. 1, pp. 1–10, Jun. 2024, doi: 10.55195/jscai.1461849.

[9] L. Siena, T. H. Saragih, R. A. Nugroho, D. Kartini, Muliadi, and W. Caesarendra, "Evaluation of the Impact of SMOTEENN on Monkeypox Case Classification Performance Using Boosting Algorithms," *Indones. J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 2, pp. 203–220, Apr. 2025, doi: 10.35882/nrgqsz63.

[10] S. Nagro, "A stacked ensemble approach for symptom-based monkeypox diagnosis," *Comput. Biol. Med.*, vol. 191, no. March, p. 110140, Jun. 2025, doi: 10.1016/j.compbiomed.2025.110140.

[11] R. Mahmood, J. Lucas, J. M. Alvarez, S. Fidler, and M. T. Law, "Optimizing Data Collection for Machine Learning," *Adv. Neural Inf. Process. Syst.*, vol. 35, no. NeurIPS, pp. 1–14, Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.01234

[12] Kaggle, "Monkeypox Patients Dataset." Accessed: Jan. 09, 2025. [Online]. Available: https://www.kaggle.com/datasets/muhammad4hmed/monkeypox-patients-dataset

[13] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian, and A. Alothaim, "Exploratory Data Analysis and Classification of a New Arabic Online Extremism Dataset," *IEEE Access*, vol. 9, pp. 161613–161626, 2021, doi: 10.1109/ACCESS.2021.3132651.

[14] E. Ibrahimi *et al.*, "Overview of data preprocessing for machine learning applications in human microbiome research," *Front. Microbiol.*, vol. 14, no. October, pp. 1–8, Oct. 2023, doi: 10.3389/fmicb.2023.1250909.

[15] E. Poslavskaya and A. Korolev, "Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding?," Dec. 2023, doi: https://doi.org/10.48550/arXiv.2312.16930.

[16] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," *IEEE Access*, vol. 9, pp. 114381–114391, 2021, doi: 10.1109/ACCESS.2021.3104357.

[17] N. Hidayat, "Improving the Accuracy of the Logistic Regression Algorithm Model using SelectKBest in Customer Prediction Based on Purchasing Behavior Patterns," vol. 1, no. 1, pp. 9–17, 2023.

[18] S. Julkaew, T. Wongsirichot, K. Damkliang, and P. Sangthawan, "Improving accuracy of vascular access quality classification in hemodialysis patients using deep learning with K highest score feature selection," *J. Int. Med. Res.*, vol. 52, no. 4, Apr. 2024, doi: 10.1177/03000605241232519.

[19] F. Gurcan and A. Soylu, "Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis," *Cancers (Basel).*, vol. 16, no. 19, p. 3417, Oct. 2024, doi: 10.3390/cancers16193417.

[20] F. Yang, K. Wang, L. Sun, M. Zhai, J. Song, and H. Wang, "A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 344, Dec. 2022, doi: 10.1186/s12911-022-02075-2.

[21] L. Strani, M. Cocchi, D. Tanzilli, A. Biancolillo, F. Marini, and R. Vitale, "One class classification (class modelling): State of the art and perspectives," *TrAC Trends Anal. Chem.*, vol. 183, no. May 2024, p. 118117, Feb. 2025, doi: 10.1016/j.trac.2024.118117.

[22] K. A. A. W. Wardana and A. M. A. Rahim, "Analisis Perbandingan Algoritma XGBoost Dan Algoritma Random Forest Untuk Klasifikasi Data Kesehatan Mental," *Log. J. Ilmu Komput. dan Pendidik.*, vol. 2, pp. 808–818, Aug. 2024.

[23] M. Maharina, "Machine Learning Models for Predicting Flood Events Using Weather Data: An Evaluation of Logistic Regression, LightGBM, and XGBoost," *J. Appl. Data Sci.*, vol. 6, no. 1, pp. 496–507, Jan. 2024, doi: 10.47738/jads.v6i1.503.

[24] S. Wu and S. Meng, "Applied Mathematics and Nonlinear Sciences (aop) (aop) Applied Mathematics and Nonlinear Sciences A Modern Communication Path for Traditional Chinese Cultural Design Concepts Based on AdaBoost Model," 2023, doi: 10.2478/10.2478/amns.2023.2.00068.

[25] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.